A MLR Approach For Medical Expenditure Prediction Model

Qiyi Zhang

University of Toronto

1 Introduction

People's lifespan has been extended in recent years because of the developing technology and better environment. However, disease is still inevitable to most of human-beings, needless to say, so is medical expense. Nowadays, especially in America, people find it almost impossible to pay for medical expenditure without insurance, and they are worried if fee from hospitals will take over their life one day. My **goal** is to find how potential medical expense will vary based on different predictors in linear regression model. I believe that this research would help me and other people get to know about the potential incoming medical expense so that we can prepare the money in advance before it is too late.

After performing background search, there are around 3100 articles about the topic of "medical expenditure prediction" on UofT online library. There is an interesting research (Journey of Women's Health, 2010) shows that patient gender were significant predictors of healthcare use and charges, and this research gives me an insight to include gender as one categorical variable in my research. Similarly, one of the article (Savings Needed to Fund Health Insurance and Health Care Expenses in Retirement: Findings from a Simulation Model, 2008) provides estimates for savings needed to cover health insurance for medical expenses for different ages of people after retirement. The result of this article is very related to my research, but I am going to consider more predictors than age of retired people. Last but not least, from the article(Feature Selection for Health Care Costs Prediction Using Weighted Evidential Regression, 2020), the researcher predicts health care costs for each patient by evaluating the data provided by Tsuyama Chuo hospital. This is also similar to what I want to work on, except that I will use the data provided by the insurance company instead of hospital, that is, I may not include variables such as patients' diagnosis record in my analysis. Therefore, my result could be biased for omitting extra information as this research did.

2 Method

In order to proceed to the prediction model of people's medial expenditure and qualify its accuracy. We will do the following steps:

2.1 Variable Selection

(1) Firstly, I cleaned out missing values of data set because it will reduce the accuracy of the model.
(2) Afterwards, I plotted histograms as well as scatterplots regarding to the the relationship between response and each variable in the data set. By referring to the literature review and common sense of my research topic, I will further prune insignificant variables from data set and proposed the potential predictors that can fit the model.

Variable	Description
age	age of primary beneficiary
sex	insurance contractor gender, female, male
bmi	Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, (objective index of body weight (kg / m ˆ 2) using the ratio of height to weight, ideally 18.5 to 24.9)
children	Number of children covered by health insurance / Number of dependents
smoker	Smoke or not
region	the beneficiary's residential area in the US, northeast, southeast, southwest, northwest

(3) I randomly chose 1000 samples out of 3630 observations from my data set, and in the ratio of 5:5, the chosen samples were further split into test and train data.
(4) Use multiple linear regression to fit a model based on the proposed variables in step (2). After that, I will discard the predictors with p-value larger than 0.05 and use remaining variables to fit a new model recursively until all predictors in the model have p-values smaller than 0.05.

2.2 Model Validation

(1) With the MLR model, I used partial F tests compared with the model with the another model that has less predictors from train data set. Moreover, I used VIF to find *multicollinearity* with other predictors.

(2) Another approach of model validation is by comparing models' adjusted r-squared, residual sum of squares (SSres), Akaike's Information Criterion (AIC） , Corrected AIC, (AICc) as well as Bayesian Information Criterio (BIC) to check the goodness of the model.

(3) Last but not the least, I created a new model based on the test data set with same predictors as my final model. If there exists obvious difference between models from train and test data sets, I would admit the inconsistency in my MLR model.

2.3 Model Violations and Diagnostics

(1) From the model, residuals are required to to be checked if the linearity and the independence are satisfied. Therefore, I would firstly check if residual plots is valid for showing the problems by testing if mean response is a single function of linear combination of the predictors and conditional mean of each predictor is not linear function with another predictor occurred. After that, by looking at residual plots, I would perceive there exists any cluster, curve or line to indicate the violation of independence and non-linearity. Moreover, based on the residual plots, I will also check if residuals have constant variance. Lastly, I will use the q&q plot to check if the graph presents a line for normality purpose. Note that if above conditions happened to be not met, I would go back to step 4 to form a new model or I could also admit the limitation among the data.

(2) Problematic observations are also required for attention to be detected and carefully handled. I will collect leverage points, outliers and inferential points of the model regarding to the train data set for more advanced validation of the model.

3 Result

3.1 Description of Data

(1) The data that will be analyzed was provided by the insurance company with their clients' personal information as well as their medical expenditures. The data is acquired from *www.kaggle.com.* In total, it includes 3630 observations on 7 variables.

(2) There are four numerical variables

Summary Table of the numerical variable

age	bmi	children	charges
Min. :18.00	Min. :15.96	Min. :0.000	Min. : 1122
1st Qu.:29.00	1st Qu.:26.69	1st Qu.:1.000	1st Qu.: 5655
Median :39.17	Median :30.20	Median :3.000	Median : 9444
Mean :38.89	Mean :30.63	Mean :2.504	Mean :12785
3rd Qu.:48.34	3rd Qu.:34.10	3rd Qu.:4.000	3rd Qu.:14680
Max. :64.00	Max. :53.13	Max. :5.000	Max. :63770

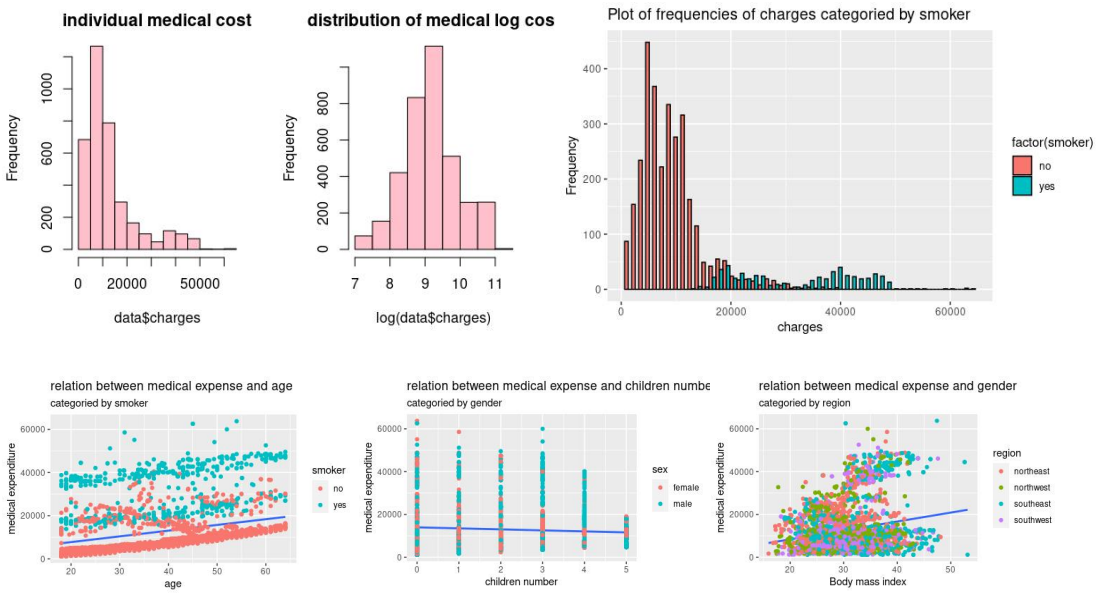
And three categorical variables:

smoker: yes/no

sex: male/female

region: southwest/northwest/southeast/northeast

(4) Result histograms and scatterplots



From the graphs above, we can derive a few important information about the model. Firstly, the data presents the fact that the log response, that is log(charges), follows normal distribution. Moreover, smokers tend to spend much more on the medical expense than non-smokers. Furthermore, it also seems that female and people with less body mass will spend less than male and people with larger body mass respectively. Last but not the least, where the people live and how many children they have seem to be insignificant variable, but because children number still vaguely imply the tendency that "more children leads less medical expense" and we still can't conclude that region is unrated to medical expense, we will still include them into our first MLR model.

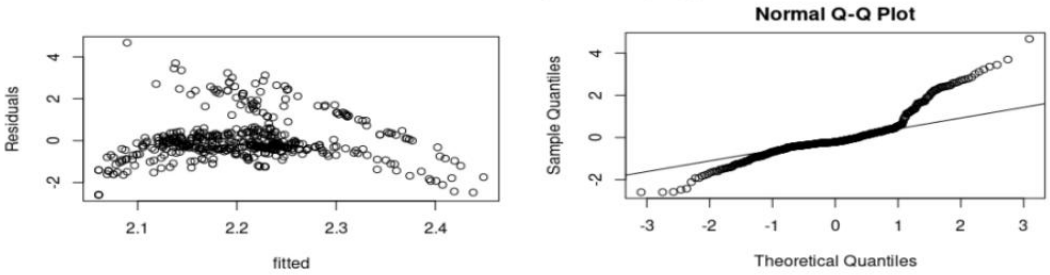
3.2 Presenting the Analysis Process and the Results

Firstly, based on the analysis of the graphs in 3.1, for model 1 will set sex, bmi, age, smoker, region, and children as predictors and the original charges as response. However, it failed conditional 1 that Y hat is not linear with Y.

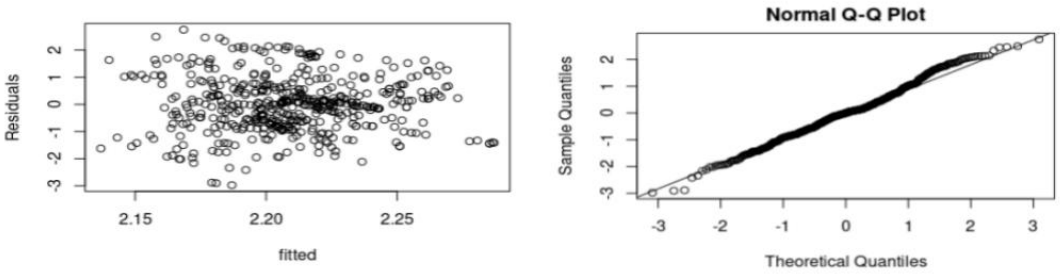
Therefore, I transform the response to log response, and kept all predictor. And it showed that bmi and sex has p value larger than 0.05 so I removed them as insignificant variables and created model 3. In model 3, it passed F tests and condition1&2, but the cluster points and curve formed on the residual plot as well as non-linearity appeared in q&q plot, hence bad MLR model. However, no matter how I transform the predictor later, the combination of remaining age, smoker, region and children can never form a model with non-linear and independent residuals.

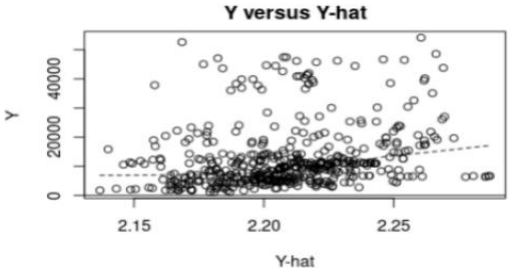
Lastly, following the common sense and experiment, I figured that children, bmi would all depend on age, and smoker should depend on sex. Then I used bmi, children, sex and region as predictor, and found that sex's p-value is larger than 0.05 hence discarded Therefore, bmi, children and region then were put into model 4, and merely passed all the requirement with some limitations that will be discussed later.

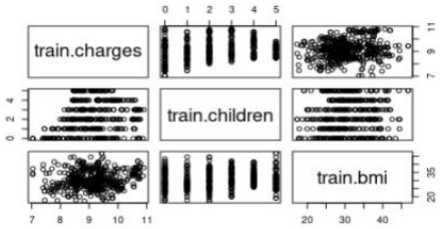
Model 3 residual plot and q&q plot



Model 4 (Final) residual plot and q&q







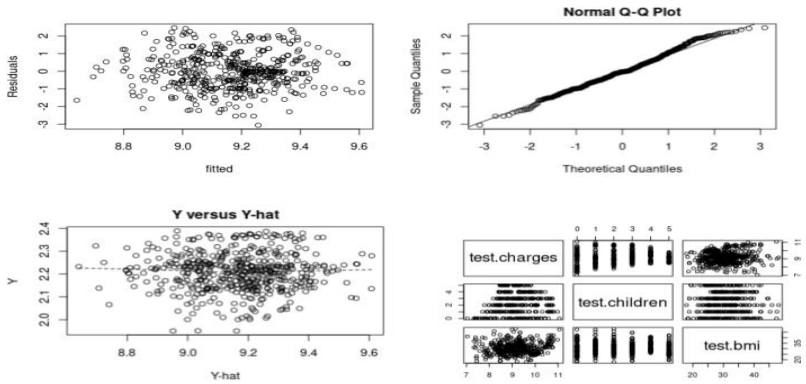
Summary Table of Model 4				
term	**estimate**	**std.error**	**statistic**	**p.value**
(Intercept)	8.625	0.190	45.386	0.000
bmi	0.020	0.006	3.460	0.001
children	0.076	0.019	4.013	0.000
regionnorthwest	-0.317	0.092	-3.441	0.001
regionsoutheast	-0.396	0.089	-4.436	0.000
regionsouthwest	-0.488	0.094	-5.190	0.000

3.3 Goodness of the Final Model

Before comparing with the test data, I firstly use partial F test with model which only contains age as predictor to compare with model 4. And the resulting p-value is smaller than 0.05, hence model 4 is better for containing more predictors.

To further test the “goodness” of the model, tests data should be used to verify the quality of model 4. Therefore, I created model 5 with bmi, children and region as

predictor but used test data this time and compared its difference with model 4. With the result graphs below, we ca see that the reproduced model is very similar to the original model derived from train data set.



Model	SSres	Rsq	Rsq_adj	AIC	AICc	BIC
Model 1	15905203719	1	1	8654	8654	8700
Model 2	73	1	1	-946	-946	-900
Model 3	1	1	1	-3148	-3148	-3110
Model 4 (Final)	264	0	0	-316	-316	-282
Model 5 (Test)	264	0	0	-309	-309	-275

VIF	bmi	children	region		problematic observations	#count
GVIF	1.009340	1.024986	1.027912		Outlier	0
					Leverage points	0
					Inferential points	0

As the table shown above, after checking with test data set, I also collected SSres, Rsq, adjusted Rsq, AIC, AICc, and BIC.From model 1 to model 4, it is sad that model 3 has larger adjusted r-squared as well as much smaller AIC, AICc, and BIC, however, model 3 should be discarded for it violates the non-linearity, independency and normality of residuals. After that, from VIF, I got very small GVIFs for each variable in the model which means that predictors are in good behaviors from multicolinearity.

Last but not least, I also checked problematic observation of the model, and it is good to know that my final model doesn't have any outliers, leverage points or inferential points which may affect the validity of the model

4 Discussion Section

4.1 Final Model Interpretation and Importance

From the final model, we can derive the conclusion that the log estimated health expense for each person can be calculated as:

Case 1: if the person lives in northwest: 8.65 + 0.02 bmi + 0.076 children number - 0.317.

Case 2: if the person lives in northeast: 8.65 + 0.02 bmi + 0.076 children number

Case 3: if the person lives in southwest: 8.65 + 0.02 bmi + 0.076 children number - 0.488

Case 4: if the person lives in south east: 8.65 + 0.02 bmi + 0.076 children number - 0.396.

To be more specific, we can conclude that each increasing body max index will lead to the addition of 0.02 in the log medial cost. Also, for each children this person has, he/she will be expected to pay 0.076 more on log medical expenditure.

In my interpretation of the final model, in order to minimize the unpleasant medical cost in the future, even though most of us can't decide where we live, but we still can control our diet and have less babies, which will also do this over-populated world a favor.

4.2 Limitations of the Analysis

To be honest, limitations of the final model is quite inevitable. We can still see cluster in the residual plot and Y hat is not quite linear with Y in condition 1 no matter how I transformed or reduced the predictors. My explanation to that is, perhaps there exist some subtle relationship between bmi, children and region that allows correlation affect the model generated. Moreover, from the conclusion of the final model, the intercept value (8.65) is still too large compared to other predictors in the model, perhaps there is a better way to acquire a more accurate model with other factors. Lastly, as we can see model 3 (with higher adjusted r-square and lower AIC) is better than model 4 even though its residuals have linearity and normality issues, hence with further study, we can improve the current model with the replacement or addition predictors.

References List

Bertakis, K. D., & Azari, R. (2010). Patient gender differences in the prediction of medical expenditures. Journal of Women's Health, 19(10), 1925-1932.

Fronstin, P., Salisbury, D., & VanDerhei, J. (2008). Savings needed to fund health insurance and health care expenses in retirement: findings from a simulation model. EBRI Issue Brief, (317).

Panay, B., Baloian, N., Pino, J. A., Peñafiel, S., Sanson, H., & Bersano, N. (2020). Feature selection for health care costs prediction using weighted evidential regression. Sensors, 20(16), 4392.